PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-128396
(43)Date of publication of application : 16.05.1997

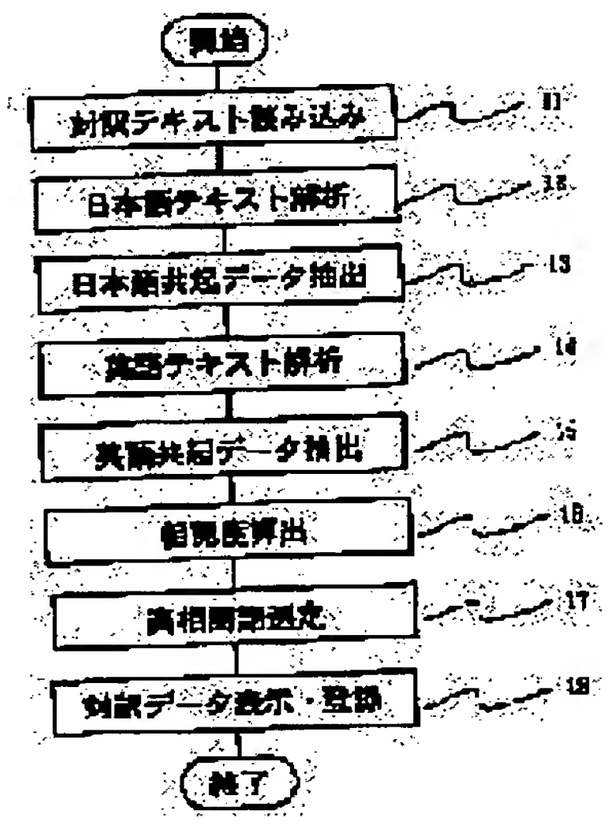
(51)Int.Cl. G06F 17/28

(21)Application number : 07-287135
(22)Date of filing : 06.11.1995

(71)Applicant : HITACHI LTD
(72)Inventor : KAJI HIROYUKI
ONO TOSHIKO

(54) PREPARATION METHOD FOR BILINGUAL DICTIONARY

(57)Abstract:
PROBLEM TO BE SOLVED: To prepare a bilingual dictionary from texts where sentences are not made to correspond by extracting bilingual data of a word from the texts and automatically preparing the bilingual dictionary.
SOLUTION: The text of a first language and the text of a second language are read (11). The words appearing in the text of the first language are extracted (12) and the set of cooccurrence words is obtained for the respective words (13). The words appearing in the text of the second language are extracted (14) and the set of the cooccurrence words on the respective words are obtained (15). The correlation degree of the set of the coocurrence words is calculated on the group of the words of the first language and the second language (16). The group of the words whose correlation degrees are largest is selected (17) and they are registered in the bilingual dictionary (18).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平9－128396

(43)公開日　平成９年(1997)５月16日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　17/28			Ｇ０６Ｆ　15/38	Ｃ

審査請求　未請求　請求項の数６　ＯＬ　（全 17 頁）

(21)出願番号　特願平7－287135

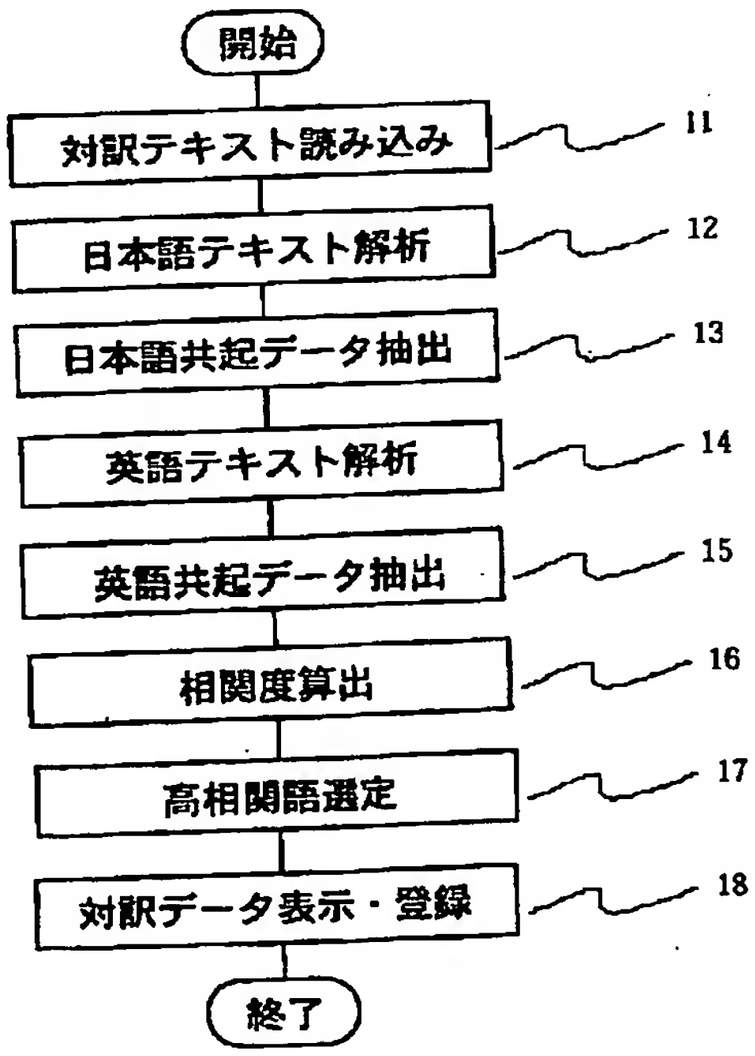
(22)出願日　平成７年(1995)11月６日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　梶　博行
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　小野　敏子
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　弁理士　中村　純之助

(54)【発明の名称】　対訳辞書作成方法

(57)【要約】
【課題】対訳テキストから語の対訳データを抽出し、対訳辞書を自動生成する。これにより、文の対応づけがなされていない対訳テキストからの辞書作成を可能とする。
【解決手段】第１言語のテキストと第２言語のテキストを読み込み（１１）、第１言語のテキストに出現する語を抽出し（１２）、各語について共起語の集合を求め（１３）、第２言語のテキストに出現する語を抽出し（１４）、各語について共起語の集合を求める（１５）。第１言語の語と第２言語の語の組の各々について共起語集合の相関度を計算し（１６）、互いに相関度が最大の語であるような語の組を選定し（１７）、対訳辞書に登録する（１８）。

図 2



1

【特許請求の範囲】
【請求項１】第１言語のテキストと第２言語のテキストを入力装置から読み込む対訳テキスト読み込みステップ、第１言語のテキスト中に出現する語を抽出する第１言語テキスト解析ステップ、第１言語のテキストに出現する語の各々についてテキスト中で共起する語の集合即ち第１の共起語集合を抽出する第１言語共起データ抽出ステップ、第２言語のテキスト中で出現する語を抽出する第２言語テキスト解析ステップ、第２言語テキストに出現する語の各々についてテキスト中で共起する語の集合即ち第２の共起語集合を抽出する第２言語共起データ抽出ステップ、第１言語の語の上記共起語集合と第２言語の語の上記共起語集合との相関度を計算する相関度算出ステップ、共起語集合の相関度に基づいて第１言語の語と第２言語の語の組を選定する高相関語選定ステップ、前記選定された語の組を対訳辞書に登録する対訳データ登録ステップから構成されることを特徴とする対訳辞書作成方法。
【請求項２】請求項１に記載の対訳辞書作成方法であって、相関度算出ステップは、対訳辞書に既登録の語の組を同一の要素とみなすことによって、第１言語の語の共起語集合と第２言語の語の共起語集合の相関度を計算することを特徴とする対訳辞書作成方法。
【請求項３】請求項１に記載の対訳辞書作成方法であって、高相関語選定ステップは、共起語集合の相関度が互いに最大の語であることを条件として、第１言語の語と第２言語の語の組を選定することを特徴とする対訳辞書作成方法。
【請求項４】請求項１に記載の対訳辞書作成方法であって、高相関語選定ステップは、対訳辞書の対訳データと語の出現頻度に基づく第２の相関度を算出し、選定する語に係わる第２の相関度より共起語集合の相関度が大きいことを選択の条件にすることを特徴とする対訳辞書作成方法。
【請求項５】請求項１に記載の対訳辞書作成方法であって、高相関語選定ステップは、共起語集合の相関度があらかじめ定めたしきい値以上であることを条件として、第１言語の語と第２言語の語の組を選定することを特徴とする対訳辞書作成方法。
【請求項６】請求項１に記載の対訳辞書作成方法であって、対訳データ登録ステップは、対訳辞書に登録する前に、語の組を表示装置に表示し、人間が登録を指示した語の組のみを対訳辞書に登録することを特徴とする対訳辞書作成方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、機械翻訳システムなどで用いられる対訳辞書の作成方法に係わり、特に対訳テキストから語の対訳データを自動的に抽出する方法に関する。

2

【０００２】
【従来の技術】機械翻訳システムにおいては対訳辞書が必須の構成要素であり、翻訳精度を向上するには，対訳辞書の語彙のカバー率を高めることが必要である。基本的な語彙の対訳辞書は、通常、機械翻訳システムのメーカが提供するが、専門用語の対訳辞書はユーザが作成することが必要であり、辞書の作成コストが問題になる。このため、対訳テキストから語の対訳データを自動的に抽出することが望まれている。専門用語の対訳辞書は、機械翻訳システムだけでなく多言語情報検索システムなどでも必須の要素であり、自動作成のニーズは非常に高い。
【０００３】対訳テキストから自動的に対訳辞書を作成する方法は、例えば、特開平７－２８８１９号に開示されている。しかし、特開平７－２８８１９号などの従来技術は、文の対応づけがなされた対訳テキストを利用するものであるという問題がある。対訳テキストは、多くの場合、テキスト全体で対訳になっているだけで、文単位での対応関係は付けられていないからである。従来技術によって対訳辞書を作成しようとすると、対訳テキストにおける文の対応づけを行う前処理が必要になる。これを人手で行うのはコスト的に問題であり、文の対応づけを自動的に行う方法も研究されている。Computational Linguistics, Vol. 19, No. 1, pp. 75-102 (1993年3月)の論文"A Program for Aligning Sentences in Bilingual Corpora"はその例である。しかし、対訳テキストには、１つの文が２つの文に対応している部分も多いし、対応する文をもたない文が含まれることさえある。従って、文の対応づけを１００％の精度で行うことは困難であり、コンピュータによる対応づけの結果を人間が確認・修正せざるを得ない。このように、文の対応づけコストを含めると、対訳辞書の作成コストの問題は解決されていないといえる。　文の対応づけを前提としない対訳テキストからの対訳辞書作成方法としては、情報処理学会自然言語処理研究会報告No. 94-12 (1993年)の論文「対訳コーパスを用いた専門用語対訳辞書の作成」がある。しかし、これは、複数の単純語から構成される複合語を抽出し、単純語の対訳辞書を参照して構成要素間の対訳関係が確認できるような複合語の組を抽出する方法であり、単純語の対訳関係の抽出には適用できない。複合語であっても、構成要素の対応関係が明白で、かつ全ての対応関係が対訳辞書に含まれる場合しか抽出できないという問題がある。
【０００４】
【発明が解決しようとする課題】本発明の目的は、上記従来技術の問題点を解決し、文の対応関係がつけられていない対訳テキストから、単純語と複合語の両方を対象として、対訳データを自動抽出する方法を提供することにある。
【０００５】

【課題を解決するための手段】上記目的を達成するため、本発明の請求項１においては、対訳関係を有する第１言語のテキストと第２言語のテキストを入力装置から読み込む対訳テキスト読み込みステップ、第１言語テキストの形態素解析を行ってテキスト中に出現する語を抽出する第１言語テキスト解析ステップ、第１言語テキストの解析結果をもとに、テキストに出現する語の各々について共起する語の集合即ち第１の共起語集合を抽出する第１言語共起データ抽出ステップ、第２言語テキストの形態素解析を行ってテキスト中に出現する語を抽出する第２言語テキスト解析ステップ、第２言語テキストの解析結果をもとに、テキストに出現する語の各々について共起する語の集合即ち第２の共起語集合を抽出する第２言語共起データ抽出ステップ、第１言語の語の第１の共起語集合と第２言語の語の第２の共起語集合との相関度を計算する相関度算出ステップ、共起語集合の相関度に基づいて第１言語の語と第２言語の語の組を選定する高相関語選定ステップ、前記選定された語の組を対訳辞書に登録する対訳データ表示・登録ステップから構成する。

【０００６】請求項２においては、上記請求項１で述べた対訳辞書作成方法の相関度算出ステップにおいて、既に対訳辞書に登録されている語の組が存在する場合は、これを同一要素と見做すことによって、第１言語の語の共起語集合と第２言語の語の共起語集合の相関度計算を行う方法としている。

【０００７】請求項３においては、上記請求項１で述べた対訳辞書作成方法の高相関語選定ステップにおいて、共起語集合の相関度が第１言語の語から見ても、第２言語の語から見ても最大となる語の組を対訳語として選定する方法としている。

【０００８】請求項４においては、上記請求項１で述べた対訳辞書作成方法の高相関語選定ステップにおいて、対訳辞書の対訳データと語自身の対訳テキスト中での出現頻度に基づいて、上記共起語集合の相関度とは異なる第２の相関度を算出し、上記共起語集合の相関度の方が第２の相関度よりも大となっている語を選定条件としている。

【０００９】請求項５においては、上記請求項１で述べた対訳辞書作成方法の高相関語選定ステップにおいて、共起語集合の相関度が予め定められた閾値以上となる第１言語の語及び第２言語の語の組を選定条件としている。

【００１０】さらに請求項６においては上記請求項１で述べた対訳辞書作成方法の対訳データ登録ステップにおいて、データ処理により得られた対訳語の組を辞書に登録する前に一度表示装置で表示し、人間が確認した後に辞書への登録を行う方法としている。

【００１１】

【発明の実施の形態】本発明の一実施例として、日英の

対訳テキストから語の対訳データを抽出する日英対訳辞書作成システムについて説明する。

【００１２】日英対訳辞書作成システムのハードウエアは、図１に示すように処理装置１、記憶装置２、入力装置３、表示装置４から構成される。処理装置１は対訳データを抽出する処理を実行する。記憶装置２は、日本語辞書２１、英語辞書２２、対訳辞書２３、日本語テキスト２４、英語テキスト２５を格納するほか、対訳データ抽出処理の作業エリア２６として用いられる。入力装置３は対訳テキストの入力に用いられ、表示装置４は抽出された対訳データの表示に用いられる。

【００１３】処理装置１が実行する対訳データ抽出処理は、図２に示すように、対訳テキスト読み込みステップ１１、日本語テキスト解析ステップ１２、日本語共起データ抽出ステップ１３、英語テキスト解析ステップ１４、英語共起データ抽出ステップ１５、相関度算出ステップ１６、高相関語選定ステップ１７、対訳データ表示・登録ステップ１８からなる。以下、各ステップについて説明する。

【００１４】（１）対訳テキスト読み込みステップ１１　対訳関係を有する日本語テキストと英語テキストを入力装置３から読み込み、記憶装置２の日本語テキスト２４と英語テキスト２５の格納エリアにそれぞれ格納する。

【００１５】（２）日本語テキスト解析ステップ１２　日本語テキスト２４を読み出して文に分割し、さらに各文を語に分割する。併せて，複数の語から構成される複合語を抽出する。

【００１６】テキストの文への分割は、テキストを構成する文字列を前方から１文字ずつチェックし、句点または改行記号が出現したら、それを文の末尾とみなすことによって行う。

【００１７】文の語への分割は、日本語辞書２１を参照して形態素解析することによって行う。形態素解析技術としては、例えば特開昭６１－４０６７１に開示されている技術を用いる。形態素解析の結果、文は語の列として表現されるが、本発明では、語の列のデータから助詞、助動詞などの機能語を除外し、名詞、動詞、形容詞、形容動詞などの内容語のみを残す処理を追加する。その理由は、機能語は言語間の対応関係が単純でなく、対訳テキストから抽出する対訳データを内容語の対訳関係に限定するのが適切であるからである。また、動詞など、活用する語はテキスト中にさまざまな変化形で出現するが、対訳辞書２３に登録されている基本形（例えば，終止形）に置き換えて出力する。

【００１８】複合語の抽出は、複合語を規定する品詞の並びを抽出することによって行う。例えば、連続する名詞の並びを複合名詞として抽出する。

【００１９】日本語テキスト解析ステップ１２によって得られる日本語テキスト解析結果２６１ａの例を図３（ａ）に示す。図３（ａ）において、「＼」は語の区切

りを、「＼＼」は文の区切りを、「＼＼＼」はテキストの終了を示す。また、語のうしろの「（m，n）」は、当該語がテキスト中の第m字で始まり第ｎ字で終わる語であることを表す。テキスト中の語の位置情報を付け加えた理由は、日本語共起データ抽出ステップ１３において、語の重なりをチェックするためである。

【００２０】（３）日本語共起データ抽出ステップ１３　図３（ａ）における日本語テキスト解析結果２６１ａをもとに、図４（ａ）における日本語出現語テーブル２６２および図５（ａ）における日本語共起頻度行列２６４を作成する。

【００２１】日本語出現語テーブル２６２は、図４（ａ）に示すように、日本語テキストに出現した語２６２１とその出現頻度２６２２を示すテーブルである。なお，図４（ａ）の日本語出現語テーブルの内容は、図３（ａ）の日本語テキスト解析結果の内容に対応している。日本語共起頻度行列２６４は、図５（ａ）に示すように、日本語出現語テーブル２６２中の語に対応する行および列からなる行列であり，（ｉ，ｊ）要素は日本語出現語テーブル２６２中の第ｉ語と第ｊ語が同一文中にそろって出現した頻度を表す。なお、図５（ａ）の日本語共起頻度行列の内容は、図３（ａ）の日本語テキスト解析結果の内容に対応している。

【００２２】日本語共起データ抽出ステップ１３の処理を図６のフローチャートに沿って説明する。

【００２３】最初に、日本語出現語テーブルのエントリ数（以後，簡単に「日本語語数」という）を表す変数を０に、日本語出現語テーブル２６２の語２６２１のフィールドを全て空白に、出現頻度２６２２のフィールドを全て０にする（１３０１）。また，図５（ａ）に示した日本語共起頻度行列２６４の全ての要素を０にする（１３０２）。さらに、日本語テキスト解析結果２６１ａ中の語を指すインデクスｉに初期値１をセットする（１３０３）。

【００２４】日本語テキスト解析結果２６１ａから第ｉ語を取り出してＸにセットし（１３０４）、Ｘを引数にして日本語出現語テーブル検索／登録サブルーチンをコールする（１３０５）。日本語出現語テーブル検索／登録サブルーチンは、引数として与えられた語が日本語出現語テーブル２６２に登録されている場合は、そのエントリ番号をリターンし、引数の語が日本語出現語テーブル２６２中に未登録である場合は，テーブル末尾に登録した上でエントリ番号をリターンするサブルーチンである。日本語出現語テーブル検索／登録サブルーチンがリターンするＸのエントリ番号をｍにセットする（１３０６）。

【００２５】次に，日本語テキスト解析結果２６１ａの中の第ｉ語のうしろの区切り記号を調べ（１３０７）。第ｉ語が文の途中の語であれば、文中のそれ以降の語を取り出し、第ｉ語との共起頻度を１増加する。そのため、まず、日本語テキスト解析結果２６１ａ中の語を指す第２のインデクスｊに初期値として（ｉ＋１）をセットする（１３０８）。日本語テキスト解析結果２６１ａから第ｊ語を取り出してＹにセットする（１３０９）。ＸとＹがテキスト中で重なっていないかどうかチェックする（１３１０）。重なっていなければ、すなわち複合語とその構成要素のような関係でなければ、Ｙを引数にして日本語出現語テーブル検索／登録サブルーチンをコールし（１３１１）、日本語出現語テーブル検索／登録サブルーチンがリターンした後、日本語出現語テーブル２６２中のＹのエントリ番号をｎにセットする（１３１２）。ｍ≠ｎであれば（１３１３），日本語共起頻度行列の（ｍ，ｎ）要素および（ｎ，ｍ）要素をそれぞれ１ずつ増加する（１３１４）。ｍ＝ｎであれば（１３１３）、（ｍ，ｎ）要素を１だけ増加する（１３１５）。第ｉ語と第ｊ語の共起を処理したあと，日本語テキスト解析結果２６１ａ中の第ｊ語のうしろの区切り記号を調べる（１３１６）。第ｊ語が文の途中の語であれば、次の語との共起データを処理するため、ｊに１を加えて（１３１７）、１３０９に戻る。第ｊ語が文末の語であれば、次の文の処理に進むため、ｉに１を加えて（１３１８）、１３０４に戻る。

【００２６】なお、日本語テキスト解析結果２６１ａの中の第ｉ語のうしろの区切り記号をチェックする１３０７において、第ｉ語がテキスト末の語であれば処理を終了する。また、第ｉ語がテキスト末以外の文末の語であれば、次の文の処理に進むため、ｉに１を加えて（１３１８）、１３０４に戻る。

【００２７】以上が日本語共起データ抽出ステップ１３の処理である。次に、日本語共起データ抽出ステップの中でコールされる日本語出現語テーブル検索／登録サブルーチンの処理を図７のフローチャートに沿って説明する。まず、引数として与えられた語をＷにセットし（１３０５１）、日本語出現語テーブルの要素を指すインデクスｋに初期値１をセットする（１３０５２）。ｋが日本語語数以下であれば（１３０５３），Ｗを日本語出現語テーブル２６２の第ｋ語と比較する（１３０５４）。一致すれば，第ｋ語の出現頻度２６２２を１だけ増加し（１３０５５），引数の語のエントリ番号としてｋをリターンする（１３０５６）。Ｗが日本語出現語テーブル２６２の第ｋ語と一致しなければ（１３０５４）、次の語と比較するため、ｋに１を加え（１３０５７）、１３０５３に戻る。ｋが日本語語数を越える時は（１３０５３）、Ｗが日本語出現語テーブル２６２に未登録であることを意味するので、日本語出現語テーブル２６２の第ｋ語としてＷを登録し（１３０５８）、日本語語数をｋに更新する（１３０５９）。このあと、第ｋ語の出現頻度２６２２を１だけ増加し（１３０５５）、引数の語のエントリ番号としてｋをリターンする（１３０５６）。

【００２８】（４）英語テキスト解析ステップ１４

英語テキスト２５を読み出して文に分割し、さらに各文を語に分割する。併せて、複数の語から構成される複合語を抽出する。

【００２９】テキストの文への分割は、テキストを構成する文字列を前方から１文字ずつチェックし、ピリオドまたは改行記号が出現したら、それを文の末尾とみなすことによって行う。なお、この方法では、「Mr.」のようにピリオドで終わる語が出現すると誤って分割される。そのような語のリストを用意し、リスト中の語に関して例外処理をすることにより、分割精度を向上させることが可能である。

【００３０】文の語への分割は、図１における英語辞書２２を参照して形態素解析することによって行う。形態素解析技術としては、例えば特開昭５８－４０６８４号の中に開示されている技術を用いる。形態素解析の結果、文は語の列として表現されるが、本発明では、語の列のデータから前置詞、冠詞、助動詞などの機能語を除外し、名詞、動詞、形容詞、副詞などの内容語のみを残す処理を追加する。また、語はテキスト中にさまざまな変化形で出現するが、図１における対訳辞書２３に登録されている基本形に置き換えて出力する。

【００３１】複合語の抽出は、複合語を規定する品詞の並びを抽出することによって行う。例えば、連続する名詞の並びや、形容詞と後接する名詞の並びを複合名詞として抽出する。

【００３２】英語テキスト解析ステップ１４によって得られる英語テキスト解析結果２６１ｂの例を図３（ｂ）に示す。英語テキスト解析結果２６１ｂに含まれる「＼」、「＼＼」、「＼＼＼」、および「（ｍ，ｎ）」の意味は日本語テキスト解析結果２６１ａにおいてと同じである。

【００３３】（５）英語共起データ抽出ステップ１５
英語テキスト解析結果２６１ｂをもとに、英語出現語テーブル２６３および英語共起頻度行列２６５を作成する。

【００３４】英語出現語テーブル２６３は、図４（ｂ）に示すように、英語テキストに出現した語２６３１とその出現頻度２６３２を示すテーブルである。なお、図４（ｂ）の英語出現語テーブルの内容は、図３（ｂ）の英語テキスト解析結果の内容に対応している。英語共起頻度行列２６５は、図５（ｂ）に示すように、英語出現語テーブル２６３中の語に対応する行および列からなる行列であり、（ｉ，ｊ）要素は英語出現語テーブル２６３中の第ｉ語と第ｊ語が同一文中にそろって出現した頻度を表す。なお、図５（ｂ）の英語共起頻度行列の内容は、図３（ｂ）の英語テキスト解析結果の内容に対応している。

【００３５】英語共起データ抽出ステップ１５の処理は、日本語共起データ抽出ステップ１３と全く同様であるので、詳細な説明は省略する。

【００３６】（６）相関度算出ステップ１６
図４（ａ）、（ｂ）に示した日本語出現語テーブル２６２、英語出現語テーブル２６３、図５（ａ）、（ｂ）に示した日本語共起頻度行列２６４、英語共起頻度行列２６５、および図１における対訳辞書２３に基づいて、図９に示す日英相関行列２６６を作成する。

【００３７】対訳辞書２３は、図８に例示するように、日本語の語２３１と英語の語２３２の組からなるレコードを記憶しており、日本語の語２３１をキーとして検索することができる。また、日英相関行列２６６は、図９に示すように、図４（ａ）に示した日本語出現語テーブル２６２中の語に対応する行、図４（ｂ）に示した英語出現語テーブル２６３中の語に対応する列からなる行列であり、（ｉ，ｊ）要素は、共起語集合に基づく、日本語出現語テーブル２６２中の第ｉ語と英語出現語テーブル２６３中の第ｊ語の相関度を表す。図９の日英相関行列の内容は、対訳辞書の内容が図８であるとの前提で、図５（ａ）の日本語共起頻度行列と図５（ｂ）の英語共起頻度行列から計算された結果である。

【００３８】日本語の語ＪＷと英語の語ＥＷの相関度Ａｓｓｏｃ（ＪＷ，ＥＷ）は次式で定義する。

【００３９】

Ａｓｓｏｃ（ＪＷ，ＥＷ）＝Ｃ／（Ａ＋Ｂ－Ｃ）

ここに、Ａ＝ＪＷの共起語集合の要素数，Ｂ＝ＥＷの共起語集合の要素数，Ｃ＝ＪＷの共起語集合とＥＷの共起語集合の積集合の要素数。

【００４０】ただし、ここでの集合は通常の集合と異なり、同一の語を複数個含むことを許し、集合の要素数は各語の個数の総和である。また、積集合は、ＪＷの共起語集合中の語とＥＷの共起語集合中の語の組が対訳辞書２３に既に登録されているとき、これらの語を同一の要素とみなし、積集合を構成する要素と考える。また、同一とみなされる語の個数が二つの集合の間で異なるとき、積集合は少ないほうの個数を含むことにする。

【００４１】例えば、図５（ａ）の日本語共起頻度行列２６４によれば「分割する」の共起語集合は次のとおりである。ここで、「／」のあとの数字が、「／」の前の語の個数を表している。

【００４２】｛日本語／１、テキスト／１、日本語テキスト／１、読み込む／１、文／２、語／１｝

また、図５（ｂ）の英語共起頻度行列２６５によれば「text」の共起語集合は次のとおりである。

【００４３】｛Japanese／１，read／１，divide／１，sentence／１｝

ここで、対訳辞書が図８に示す３つのレコードのみを含むとすれば、「分割する」の共起語集合と「text」の共起語集合の積集合は次のようになる。ここでは、日本語の語と英語の語を＝で結ぶことによって、同一とみなされた要素であることを示している。

【００４４】

9

{読み込む＝read／１，文＝sentence／１}
従って、上の定義式による「分割する」と「text」の相関度は次のようになる。
【００４５】Ａｓｓｏｃ（分割する，text）＝２／（７＋４－２）＝２／９
相関度算出ステップ１６は、図１０に示すように、日英対訳行列作成サブステップ１６ａ、日英仮想共起頻度行列計算サブステップ１６ｂ、日英相関行列計算サブステップ１６ｃの３つのサブステップにわけられる。
【００４６】日英対訳行列作成サブステップ１６ａは、図４（ａ）（ｂ）に示した日本語出現語テーブル２６２、英語出現語テーブル２６３、および対訳辞書２３から日英対訳行列２６７を作成する。日英対訳行列２６７は、図１１に示すように、日本語出現語テーブル２６２中の語に対応する行、英語出現語テーブル２６３中の語に対応する列からなる行列である。（ｉ，ｊ）要素の値は、日本語出現語テーブル２６２中の第ｉ語と英語出現語テーブル２６３中の第ｊ語の組が対訳辞書２３に含まれているとき１、対訳辞書２３に含まれていないとき０である。なお、図１１の日英対訳行列の内容は図８の対訳辞書の内容に対応している。
【００４７】日英対訳行列作成サブステップ１６ａの処理を図１２のフローチャートに沿って説明する。
【００４８】最初に、日英対訳行列２６７の全要素の値を０にする（１６０１）。次に、日本語出現語テーブル２６２の要素を指すインデクスｉに初期値１をセットし（１６０２）、ｉが日本語語数と一致するまで（１６１３）、ｉに順次１を加えながら（１６１４）、以下の処理を繰り返す。
【００４９】日本語出現語テーブル２６２中の第ｉ語（以後、簡単に「日本語の第ｉ語」という）をキーとして対訳辞書２３を検索する（１６０３）。一つ以上の訳語が得られた場合（１６０４）、訳語を指すインデクスｒに初期値１をセットし（１６０５）、ｒが得られた訳語数と一致するまで（１６１１）、順次ｒに１を加えながら（１６１２）、次の処理を繰返し実行する。英語出現語テーブル２６３の要素を指すインデクスｊに初期値１をセットし（１６０６）、ｊが英語出現語テーブルのエントリ数（以後、簡単に「英語語数」という）と一致するまで（１６０７）、ｊに１を加えながら（１６０９）、第ｒ訳語と英語出現語テーブル中の第ｊ語（以後、簡単に「英語の第ｊ語」という）を比較する動作（１６０８）を繰返す。一致するｊに到達すると、日英対訳行列２６７の（ｉ，ｊ）要素の値を１にする（１６１０）。
【００５０】日英仮想共起頻度行列計算サブステップ１６ｂは、日本語共起頻度行列２６４（図５（ａ））と日英対訳行列２６７（図１１）から日英仮想共起頻度行列２６８（図１３）を計算する。日英仮想共起頻度行列２６８は、図１３に示すように、日本語出現語テーブル２６２中の語に対応する行、英語出現語テーブル２６３中の語に対応する列からなる行列であり、（ｉ，ｊ）要素は日本語出現語テーブル２６２中の第ｉ語と英語出現語テーブル２６３中の第ｊ語との仮想的な共起頻度を表す。「仮想的な」共起とは、日本語テキストにおいて二つの語ＪＷ１とＪＷ２が共起する場合、ＪＷ２の英訳語がＪＷ１と共起するとみなすことを意味する。なお、図１３の日英仮想共起頻度行列の内容は、図５（ａ）の日本語共起頻度行列と図１１の日英対訳行列から計算された内容である。

10

【００５１】日英仮想共起頻度行列計算サブステップ１６ｂの処理を図１４のフローチャートに沿って説明する。
【００５２】日本語出現語テーブル２６２の要素を指すインデクスｉに初期値１をセットし（１６２１）、ｉが日本語語数と一致するまで（１６３０）順次ｉに１を加えながら（１６３１）以下の処理を行い、さらにその過程において英語出現語テーブル２６３の要素を指すインデクスｊに初期値１をセットし（１６２２）、ｊが英語語数と一致するまで（１６２８）ｊに１を加える操作を実行しつつ（１６２９）以下の処理を繰り返す。
【００５３】日英仮想共起頻度行列２６８の要素の値を計算するための変数Ｘに初期値０をセットし、日本語出現語テーブル２６２の要素を指す第２のインデクスｋに初期値１をセットする（１６２３）。日本語共起頻度行列２６４の（ｉ，ｋ）要素と日英対訳行列２６７の（ｋ，ｊ）要素の積をＸに加算する（１６２４）。ｋが日本語語数より小さければ（１６２５）、ｋに１を加え（１６２６）、１６２４に戻る。ｋが日本語語数に等しいならば、その時点のＸの値を日英仮想共起頻度行列２６８の（ｉ，ｊ）要素の値として出力する（１６２７）。
【００５４】日英相関行列計算サブステップ１６ｃ（図１０）は、日本語共起頻度行列２６４，日英仮想共起頻度行列２６８と英語共起頻度行列２６５から日英相関行列２６６を計算する。このステップの処理を図１５のフローチャートに沿って説明する。
【００５５】日本語出現語テーブル２６２の要素を指すインデクスｉに初期値１をセットし（１６４１）、ｉが日本語語数と一致するまで（１６５５）、順次ｉに１を加えながら（１６５６）以下の処理を繰り返す。
【００５６】日本語の第ｉ語の共起語集合の要素数を累計する変数Ａに初期値０をセットし、日本語出現語テーブル２６２の要素を指す第２のインデクスｋに初期値１をセットする（１６４２）。日本語共起頻度行列２６４の（ｉ，ｋ）要素をＡに加える（１６４３）。ｋが日本語語数より小さければ（１６４４）、ｋに１を加え（１６４５）、１６４３に戻る。ｋが日本語語数に等しければ、その時点のＡの値が、日本語の第ｉ語の共起語集合の要素数を表している。

【００５７】英語出現語テーブル２６３の要素を指すインデクスｊに初期値１をセットし（１６４６）、ｊが英語語数と一致するまで（１６５３）順次ｊに１を加えながら（１６５４）以下の処理を繰り返す。

【００５８】英語の第ｊ語の共起語集合の要素数を累計する変数Ｂ、および日本語の第ｉ語の共起語集合と英語の第ｊ語の共起語集合の積集合の要素数を累計する変数Ｃに初期値０をセットし、英語出現語テーブル２６３の要素を指す第２のインデクスｋに初期値１をセットする（１６４７）。英語共起頻度行列２６５の（ｊ，ｋ）要素をＢに加える（１６４８）。また、日英仮想共起頻度行列２６８の（ｉ，ｋ）要素と英語共起頻度行列２６５の（ｊ，ｋ）要素の最小値をＣに加える（１６４９）。ｋが英語語数より小さければ（１６５０）、ｋに１を加え（１６５１）、１６４８に戻る。ｋが英語語数に等しければ、その時点のＢの値が、英語の第ｊ語の共起語集合の要素数を表し、その時点のＣの値が、日本語の第ｉ語の共起語集合と英語の第ｊ語の共起語集合の積集合の要素数を表している。以上のようにして得られたＡ、Ｂ、Ｃの値からＣ／（Ａ＋Ｂ－Ｃ）を計算し、日英相関行列２６６の（ｉ，ｊ）要素として出力する（１６５２）。

【００５９】（７）高相関語選定ステップ１７

日英相関行列２６６および日英対訳行列２６７に基づいて、対訳データ２６９を抽出する。抽出される対訳データ２６９は、図１６に例示するように、日本語の語２６９１、英語の語２６９２と相関度２６９３の組である。図１６の対訳データは、図９の日英相関行列と図１１の日英対訳行列から得られた対訳データである。

【００６０】高相関語選定ステップ１７の処理を図１７、図１８のフローチャートに沿って説明する。なお、図１７と図１８はＬ１、Ｌ２、Ｌ３でそれぞれ接続されている。

【００６１】日本語出現語テーブル２６２の要素を指すインデクスｉに初期値１をセットし（１７０１）、ｉが日本語語数と一致するまで（１７２６）順次ｉに１を加えながら（１７２７）、以下の処理を繰り返す。

【００６２】英語出現語テーブル２６３の要素を指すインデクスｊに初期値１をセットし、日本語出現語テーブルの第ｉ語（以後、簡単に「日本語の第ｉ語」という）に係わる相関度の最大値を記憶する変数Ａの初期値を－１にする（１７０２）。日本語の第ｉ語と英語出現語テーブルの第ｊ語（以後、簡単に「英語の第ｊ語」という）との相関度をＡと比較し（１７０３）、Ａより大であればＡをその値に更新し、その時点のｊの値をｊｍにセーブし、日本語の第ｉ語との相関度がＡである英語の語数を記憶する変数Ｎの値を１にする（１７０４）。日本語の第ｉ語と英語の第ｊ語との相関度がＡと等しければ、Ｎを更新する（１７０５）。ｊが英語語数より小であれば（１７０６）、英語の次の語について処理するため、ｊに１を加え（１７０７）、１７０３に戻る。

【００６３】英語出現語テーブルの全ての語について上記の処理が終了したら、日本語の第ｉ語との相関度が最大値Ａである英語の語数Ｎが１かどうかチェックする（１７０８）。Ｎが１でなければ、日本語の第ｉ語に係わる対訳データは抽出できなかったと判断し、１７２６に飛ぶ。Ｎが１であれば、日本語の第ｉ語と英語の第ｊｍ語の組が対訳の候補になるので、英語の第ｊｍ語に係わる相関度とＡとの大小をチェックする。すなわち、日本語出現語テーブル２６２の要素を指す第２のインデクスｋに初期値１をセットし（１７０９）、ｋが日本語語数と一致するまで（１７１１）、ｋに１を加えながら（１７１２）、日本語の第ｋ語と英語の第ｊｍ語の相関度をＡと比較する（１７１０）。英語の第ｊｍ語との相関度がＡより大きな日本語の語があれば、日本語の第ｉ語と英語の第ｊｍ語の組は対訳でないと判断し、１７２６に飛ぶ。英語の第ｊｍ語との相関度がＡより大きな日本語の語がなければ、日本語の第ｉ語と英語の第ｊｍ語の組は、どちらの語からみても最大の相関度であるので、対訳の候補として残す。

【００６４】次に、日本語の第ｉ語と英語の第ｊｍ語の相関度Ａを、日本語の第ｉ語に係わる直接相関度、および英語の第ｊｍ語に係わる直接相関度と比較する。ここで、直接相関度とは、共起語集合に基づく相関度と異なり、対訳辞書２３に対訳として登録されているかどうかということと、語自身の対訳テキスト中での出現頻度に基づく相関度である。

【００６５】英語出現語テーブル２６３の要素を指すインデクスｊに初期値１をセットし（１７１３）、英語語数と一致するまで（１７１７）順次ｊに１を加えながら（１７１８）、日本語の第ｉ語と英語の第ｊ語の直接相関度ＢをＡと比較する。すなわち、日英対訳行列２６７の（ｉ，ｊ）要素が１であれば（１７１４）、日本語の第ｉ語の出現頻度２６２２と英語の第ｊ語の出現頻度２６３２の最小値を最大値で除した値をＢとし（１７１５）、ＡとＢとの大小を比較する（１７１６）。日本語の第ｉ語との直接相関度ＢがＡ以上である英語の語があれば、日本語の第ｉ語と英語の第ｊｍ語の組は対訳でないと判断し、１７２６に飛ぶ。

【００６６】同様に、日本語出現語テーブル２６２の要素を指す第２のインデクスｋに初期値１をセットし（１７１９）、ｋが日本語語数と一致するまで（１７２３）、ｋに１を加えながら（１７２４）、日本語の第ｋ語と英語の第ｊｍ語の直接相関度ＢをＡと比較する。すなわち、日英対訳行列２６７の（ｋ，ｊｍ）要素が１であれば（１７２０）、日本語の第ｋ語の出現頻度２６２２と英語の第ｊｍ語の出現頻度２６３２の最小値を最大値で除した値をＢとし（１７２１）、ＡとＢとの大小を比較する（１７２２）。英語の第ｊｍ語との直接相関度ＢがＡ以上である日本語の語があれば、日本語の第ｉ語

と英語の第ｊｍ語の組は対訳でないと判断し、１７２６に飛ぶ。

【００６７】日本語の第ｉ語あるいは英語の第ｊｍ語に関し、Ａより大きな直接相関度をもつ語がない場合には、日本語の第ｉ語、英語の第ｊｍ語、および相関度Ａの組を対訳データ２６９として出力する（１７２５）。

【００６８】（８）対訳データ表示・登録ステップ１８　高相関語選定ステップ１７で選定された対訳データ２６９を表示装置４に表示する。ユーザは、表示された対訳データの各々について、対訳辞書２３に登録するか否かを入力装置３から指示することができる。対訳データの登録が指示されると、当該対訳データの日本語の語２６９１と英語の語２６９２を組にして対訳辞書２３に登録する。

【００６９】以上の（１）から（８）のステップを実行することにより、対訳テキストから語の対訳データを抽出し、対訳辞書を充実させていくことができる。例えば、図８に示す３つの対訳データから成る対訳辞書を利用して、図３（ａ）と図３（ｂ）に示す対訳テキストを処理することにより、図１６に示す２つの対訳データが対訳辞書に追加される。

【００７０】以上、説明したのは一実施例であり、各ステップに種々のバリエーションを考えることができる。

【００７１】「共起する語」について、上記実施例では「同一文中に出現する語」としたが、大きさｎのウインドゥに含まれる語を採用してもよい。例えば、大きさ７のウインドゥの場合、ある語と共起する語とは、その語の前後それぞれ３語の範囲に出現する語である。また、構文的に関係のある語（修飾／被修飾の関係にある語）を共起する語とする方法も考えられる。

【００７２】共起データ抽出ステップにおいて、上記実施例では、複合語とその構成語の関係は共起関係ではないので、複合語とその構成語の組は共起頻度としてはカウントしていない。しかし、複合語と構成語は、共起とは違った意味で関連がある。すなわち、構成要素の間の対訳関係は、複合語の対訳関係抽出の手掛かりになる。従って、複合語の共起語の中にその構成語を含めて相関度を計算する方法も考えられる。この場合、相関度は、共起語の対訳知識だけでなく構成語の対訳知識を反映したものになる。

【００７３】共起語集合の相関度についても、上記実施例以外に種々の定義が可能である。例えば、共起特性をベクトルで表現し、ベクトル間の角度が小さいほど相関が高いとする方法が考えられる。ここで、日本語の語の共起特性は、各成分が日本語の語に対応するベクトルで表現され、一方、英語の語の共起特性は、各成分が英語の語に対応するベクトルで表現される。従って、語の対訳関係に基づいて成分を対応づけた上でベクトル間の角度を計算することにする。

【００７４】高相関語選択ステップにおいて、上記実施

例では、相関度が最大の語の組であっても、より大きな直接相関度をもつ語が存在する場合は除外している。この処理において、直接相関度に適当な重みをつけることが考えられる。また、直接相関度との比較処理を完全に省略することも考えられる。さらに、簡単に、あらかじめ定めたしきい値以上の相関度をもつ全ての語の組を選定する方法も考えられる。

【００７５】

【発明の効果】本発明によれば，対訳辞書に既登録の対訳データを利用して、対訳辞書に未登録の対訳データを対訳テキストから自動的に抽出し、対訳辞書を充実させていくことができる。従来技術と異なり、文の対応がつけられていない対訳テキストから対訳データを抽出できることが本発明の顕著な効果である。

【図面の簡単な説明】

【図１】日英対訳辞書作成システムのハードウエア構成図。

【図２】対訳データ抽出処理のフローチャート。

【図３】（ａ）は日本語テキスト解析結果の例を示す図、（ｂ）は英語テキスト解析結果の例を示す図。

【図４】（ａ）は日本語出現語テーブルの例を示す図、（ｂ）は英語出現語テーブルの例を示す図。

【図５】（ａ）は日本語共起頻度行列の例を示す図、（ｂ）は英語共起頻度行列の例を示す図。

【図６】日本語共起データ抽出処理のフローチャート。

【図７】日本語出現語テーブル検索／登録サブルーチンのフローチャート。

【図８】対訳辞書の例を示す図。

【図９】日英相関行列の例を示す図。

【図１０】相関度算出処理のフローチャート。

【図１１】日英対訳行列の例を示す図。

【図１２】日英対訳行列作成処理のフローチャート。

【図１３】日英仮想共起頻度行列の例を示す図。

【図１４】日英仮想共起頻度行列計算処理のフローチャート。

【図１５】日英相関行列計算処理のフローチャート。

【図１６】抽出された対訳データの例を示す図。

【図１７】高相関語選定処理のフローチャート（その１）。

【図１８】高相関語選定処理のフローチャート（その２）。

【符号の説明】

１　処理装置　　２　記憶装置

３　入力装置　　４　表示装置

１１　対訳テキスト読み込みステップ　　１２　日本語テキスト解析ステップ

１３　日本語共起データ抽出ステップ　　１４　英語テキスト解析ステップ

１５　英語共起データ抽出ステップ　　１６　相関度算出ステップ

15

１７　高相関語選定ステップ　　１８　対訳データ表示・登録ステップ
２１　日本語辞書　　２２　英語辞書
２３　対訳辞書　　２４　日本語テキスト
２５　英語テキスト　　２６　作業エリア
２６１ａ　日本語テキスト解析結果　　２６１ｂ　英語テキスト解析結果

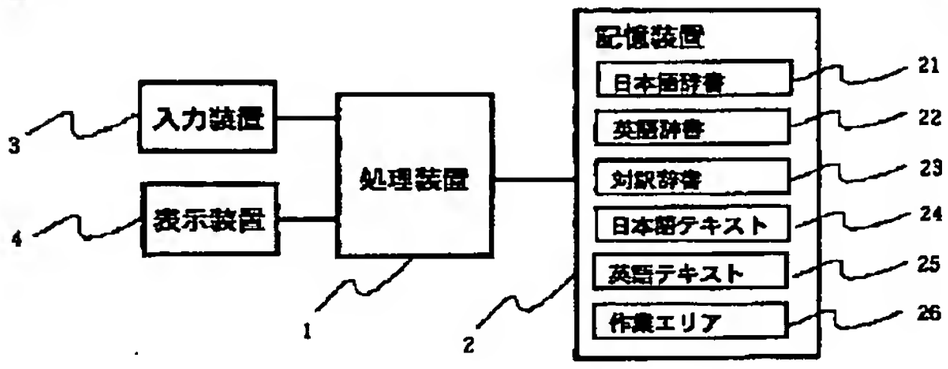
【図１】

16

２６２　日本語出現語テーブル　　２６３　英語出現語テーブル
２６４　日本語共起頻度行列　　２６５　英語共起頻度行列
２６６　日英相関行列　　２６７　日英対訳行列
２６８　日英仮想共起頻度行列　　２６９　抽出された対訳データ

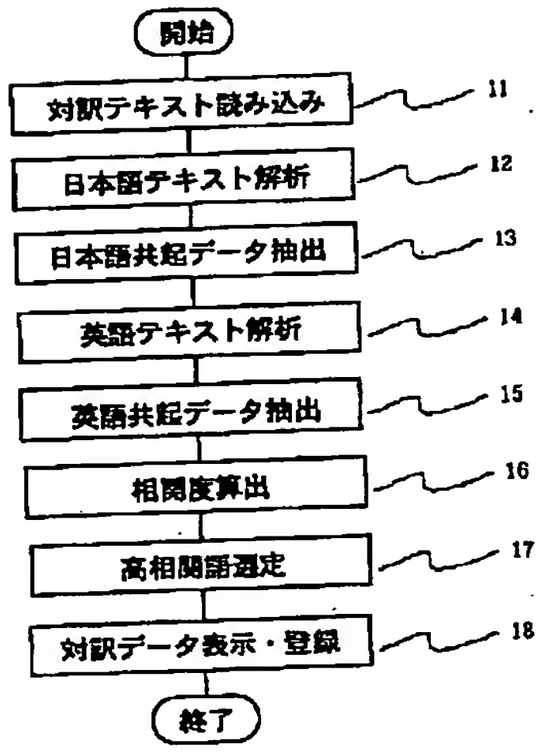
【図２】

図 1



【図３】

図 2



【図４】

図 3

（ａ）

入力テキスト

日本語テキストを読み込み，文に分割する．次に，各文を語に分割する．

261a

解析結果

日本語(1,3)＼テキスト(4,7)＼日本語テキスト(1,7)＼読み込む(9,12)＼文(14,14)＼分割する(16,19)＼＼文(25,25)＼語(27,27)＼分割する(29,32)＼＼＼

（ｂ）

入力テキスト

A Japanese text is read and divided into sentences. Then, each sentence is divided into words.

261b

解析結果

Japanese(3,10)＼text(12,15)＼Japanese text(3,15)＼read(20,23)＼divide(29,35)＼sentence(42,50)＼＼sentence(65,72)＼divide(77,83)＼word(90,94)＼＼＼

図 4

（ａ）

262

2621	2622
日本語	1
テキスト	1
日本語テキスト	1
読み込む	1
文	2
分割する	2
語	1

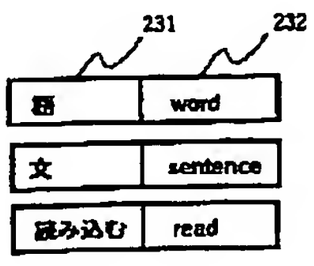
（ｂ）

263

2631	2632
Japanese	1
text	1
Japanese text	1
read	1
divide	2
sentence	2
word	1

【図８】

図 8



【図５】

図５

（ａ）

	日本語テキスト	テキスト	日本語テキスト	読み込む	文	分割する	語
日本語テキスト	0	1	0	1	1	1	0
テキスト	1	0	0	1	1	1	0
日本語テキスト	0	0	0	1	1	1	0
読み込む	1	1	1	0	1	1	0
文	1	1	1	1	0	2	1
分割する	1	1	1	1	2	0	1
語	0	0	0	0	1	1	0

264

（ｂ）

	Japanese text	Japanese text	read	divide	sentence	word
Japanese text	0	1	0	1	1	1	0
text	1	0	0	1	1	1	0
Japanese text	0	0	0	1	1	1	0
read	1	1	1	0	1	1	0
divide	1	1	1	1	0	2	1
sentence	1	1	1	1	2	0	1
word	0	0	0	0	1	1	0

265

【図９】

図９

	Japanese text	Japanese text	read	divide	sentence	word
日本語テキスト	1/3	1/3	2/5	1/8	2/9	1/10	1/5
テキスト	1/3	1/3	2/5	1/8	2/9	1/10	1/5
日本語テキスト	2/5	2/5	1/2	1/7	1/4	1/9	1/4
読み込む	1/8	1/8	1/7	1/9	1/11	0	1/6
文	1/10	1/10	1/9	0	1/6	1/6	0
分割する	2/9	2/9	1/4	1/11	2/5	1/6	1/8
語	1/5	1/5	1/4	1/6	1/8	0	1/3

266

【図１１】

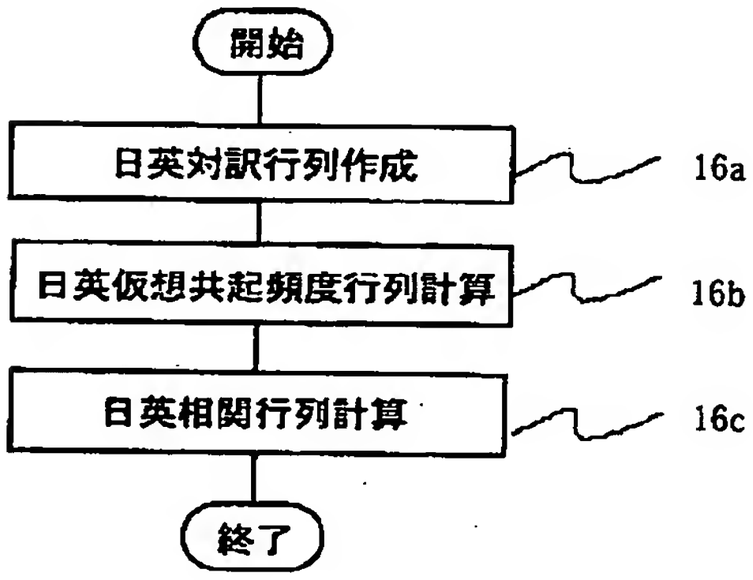
図１１

	Japanese text	Japanese text	read	divide	sentence	word
日本語テキスト	0	0	0	0	0	0	0
テキスト	0	0	0	0	0	0	0
日本語テキスト	0	0	0	0	0	0	0
読み込む	0	0	0	1	0	0	0
文	0	0	0	0	0	1	0
分割する	0	0	0	0	0	0	0
語	0	0	0	0	0	0	1

267

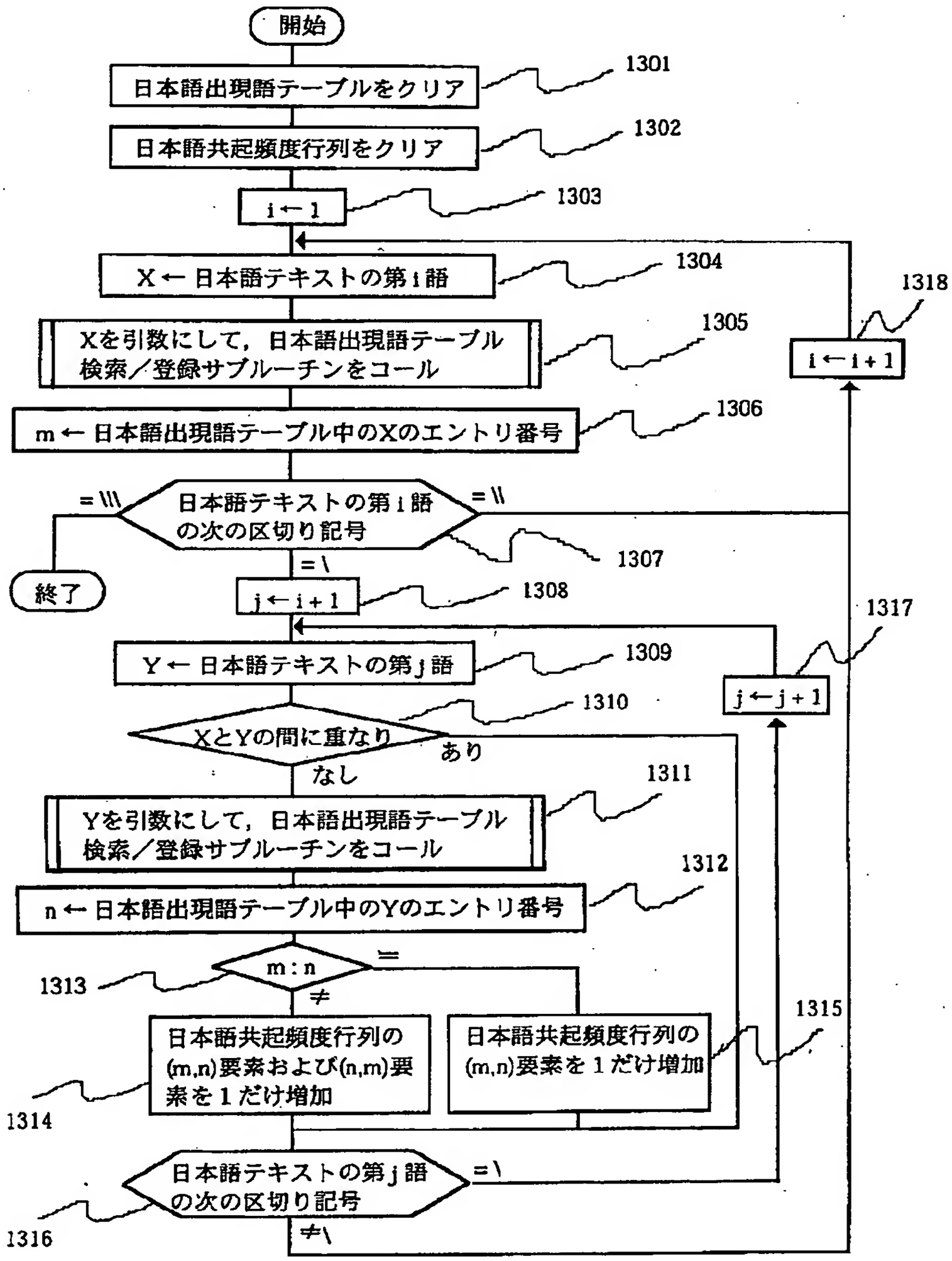
【図１０】

図１０



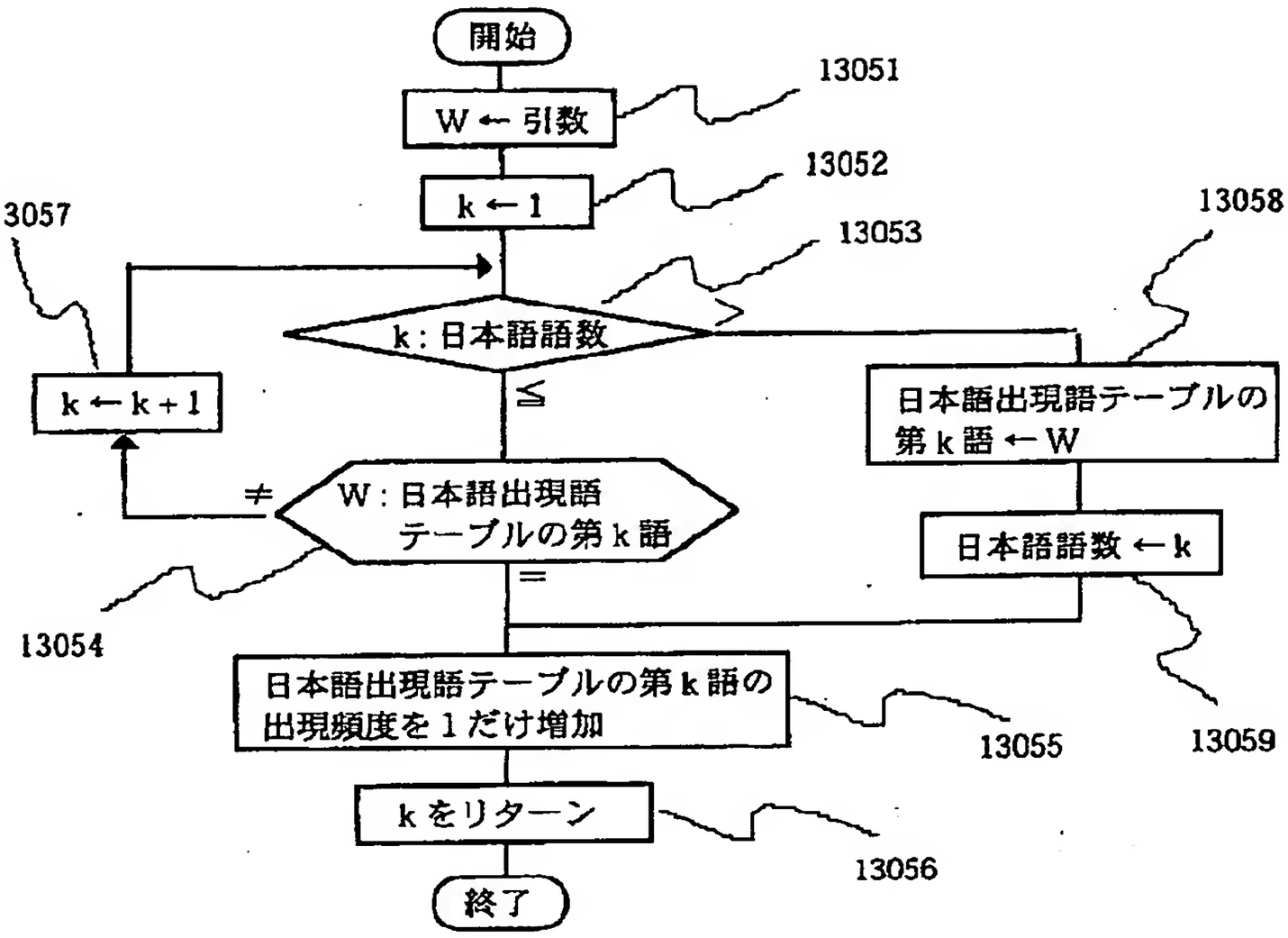
【図６】

図 6


開始
日本語出現語テーブルをクリア
1301
日本語共起頻度行列をクリア
1302
i←1
1303
X←日本語テキストの第i語
1304
Xを引数にして，日本語出現語テーブル検索／登録サブルーチンをコール
1305
m←日本語出現語テーブル中のXのエントリ番号
1306
日本語テキストの第i語の次の区切り記号
=\\\
=\\
=\
1307
終了
j←i+1
1308
Y←日本語テキストの第j語
1309
XとYの間に重なり
1310
あり
なし
Yを引数にして，日本語出現語テーブル検索／登録サブルーチンをコール
1311
n←日本語出現語テーブル中のYのエントリ番号
1312
m:n
=
≠
1313
日本語共起頻度行列の(m,n)要素および(n,m)要素を１だけ増加
1314
日本語共起頻度行列の(m,n)要素を１だけ増加
1315
日本語テキストの第j語の次の区切り記号
=\
≠\
1316
j←j+1
1317
i←i+1
1318

【図７】

図 7



【図１３】

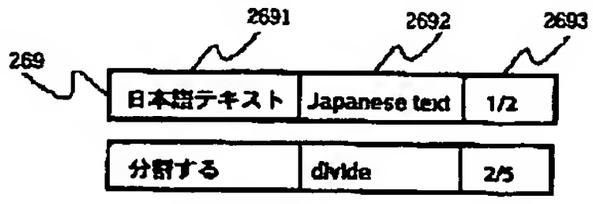
図 1 3

	Japanese	text	Japanese text	read	divide	sentence	word
日本語	0	0	0	1	0	1	0
テキスト	0	0	0	1	0	1	0
日本語テキスト	0	0	0	1	0	1	0
読み込む	0	0	0	0	0	1	0
文	0	0	0	1	0	0	1
分割する	0	0	0	1	0	2	1
語	0	0	0	0	0	1	0

258

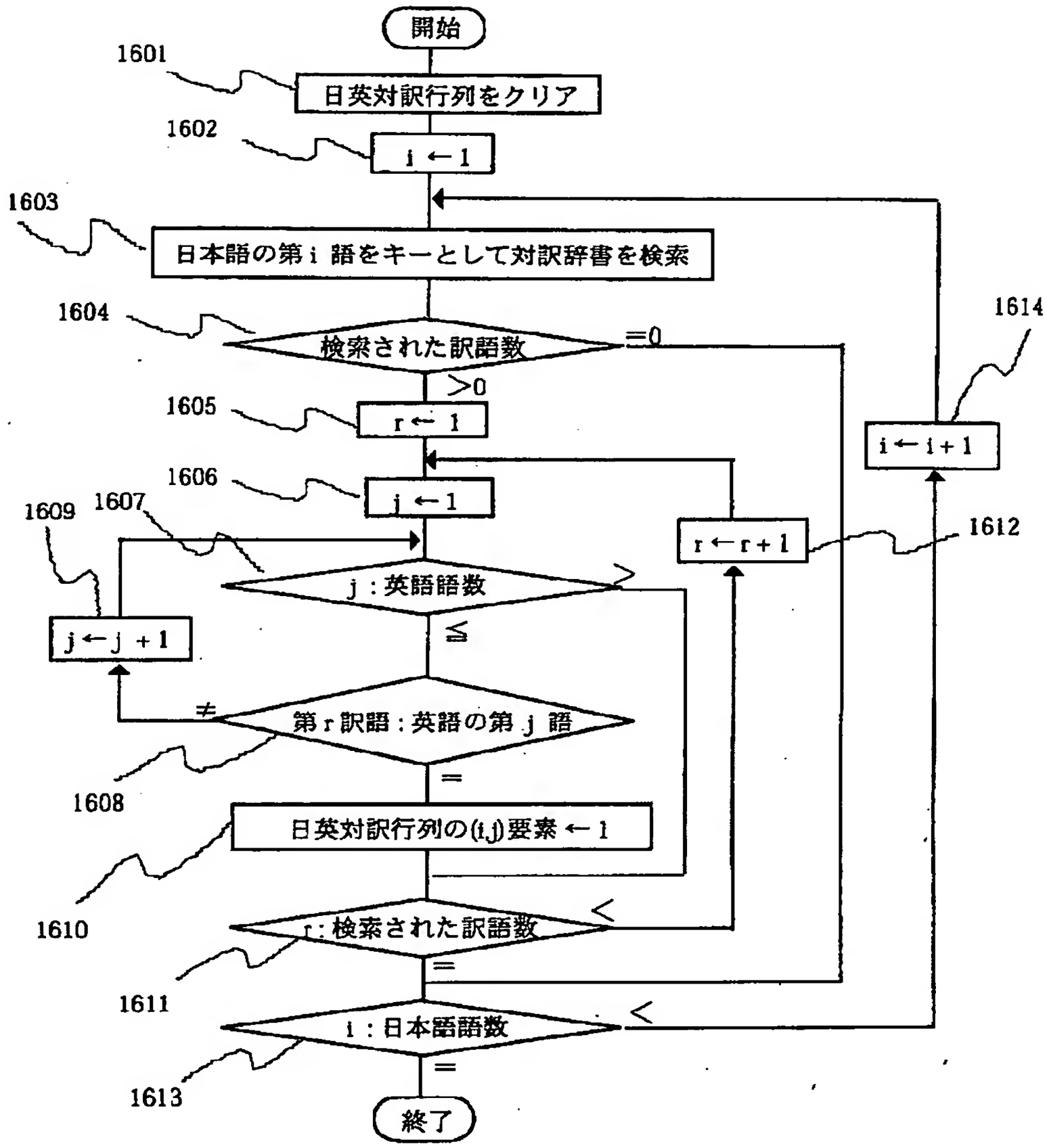
【図１６】

図 1 6



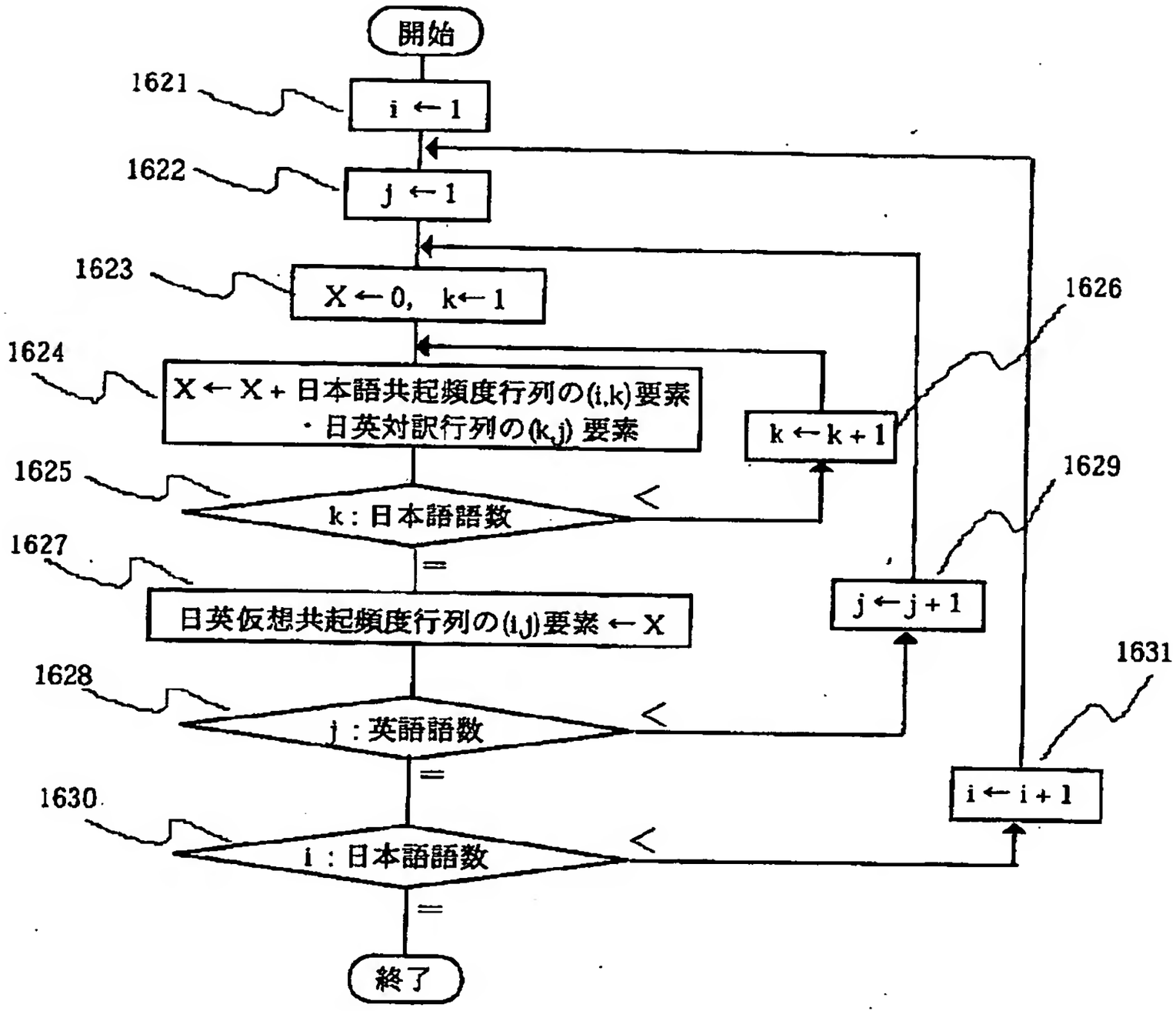
【図１２】

図１２


開始
1601
日英対訳行列をクリア
1602
i ← 1
1603
日本語の第 i 語をキーとして対訳辞書を検索
1604
検索された訳語数
=0
>0
1605
r ← 1
1606
j ← 1
1609
1607
j ：英語語数
>
≦
j ← j + 1
≠
第 r 訳語 : 英語の第 j 語
1608
=
日英対訳行列の(i,j)要素 ← 1
1610
r : 検索された訳語数
<
=
1611
i ：日本語語数
<
=
1613
終了
r ← r + 1
1612
i ← i + 1
1614

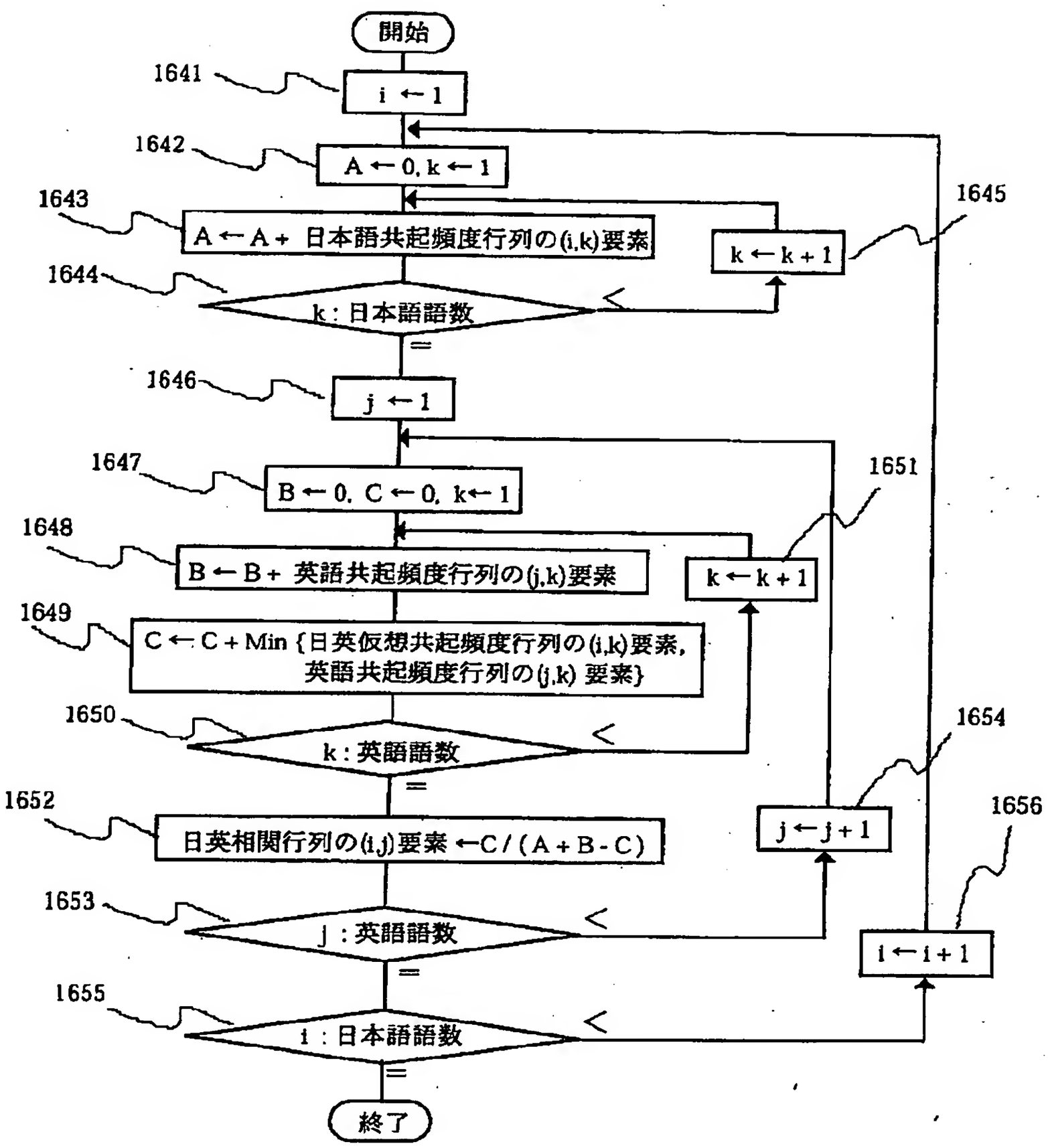
【図１４】

図１４


開始
1621
i ← 1
1622
j ← 1
1623
X ← 0, k ← 1
1624
X ← X + 日本語共起頻度行列の(i,k)要素
・日英対訳行列の(k,j) 要素
k ← k + 1
1625
k：日本語語数
<
=
1626
1627
日英仮想共起頻度行列の(i,j)要素 ← X
j ← j + 1
1628
j：英語語数
<
=
1629
1630
i：日本語語数
<
=
i ← i + 1
1631
終了

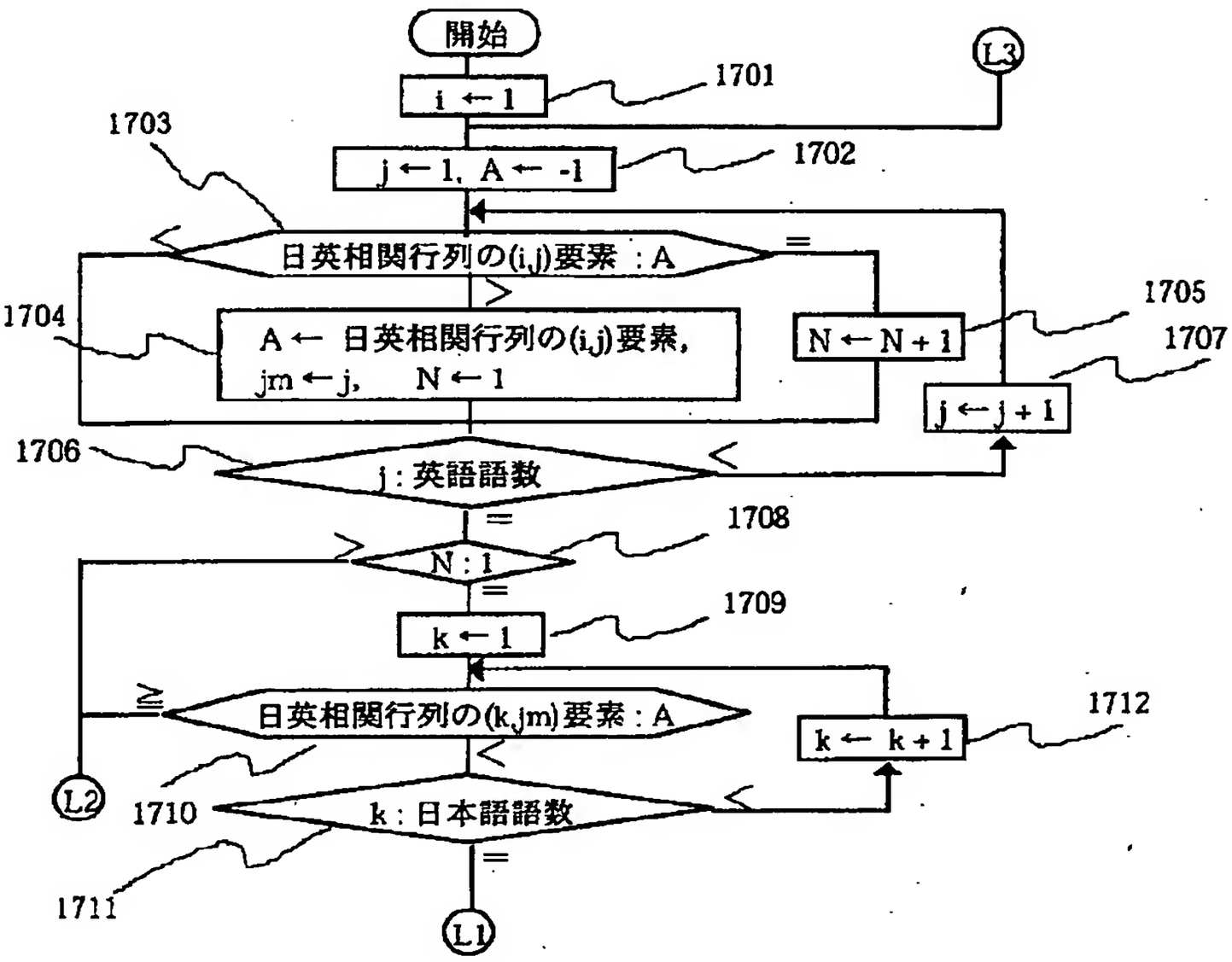
【図１５】

図１５


開始
1641
i ← 1
1642
A ← 0, k ← 1
1643
A ← A + 日本語共起頻度行列の(i,k)要素
1644
k : 日本語語数
<
=
k ← k + 1
1645
1646
j ← 1
1647
B ← 0, C ← 0, k ← 1
1651
1648
B ← B + 英語共起頻度行列の(j,k)要素
k ← k + 1
1649
C ← C + Min {日英仮想共起頻度行列の(i,k)要素,
英語共起頻度行列の(j,k) 要素}
1650
k : 英語語数
<
=
1654
1652
日英相関行列の(i,j)要素 ←C / (A + B - C)
j ← j + 1
1656
1653
j : 英語語数
<
=
i ← i + 1
1655
i : 日本語語数
<
=
終了

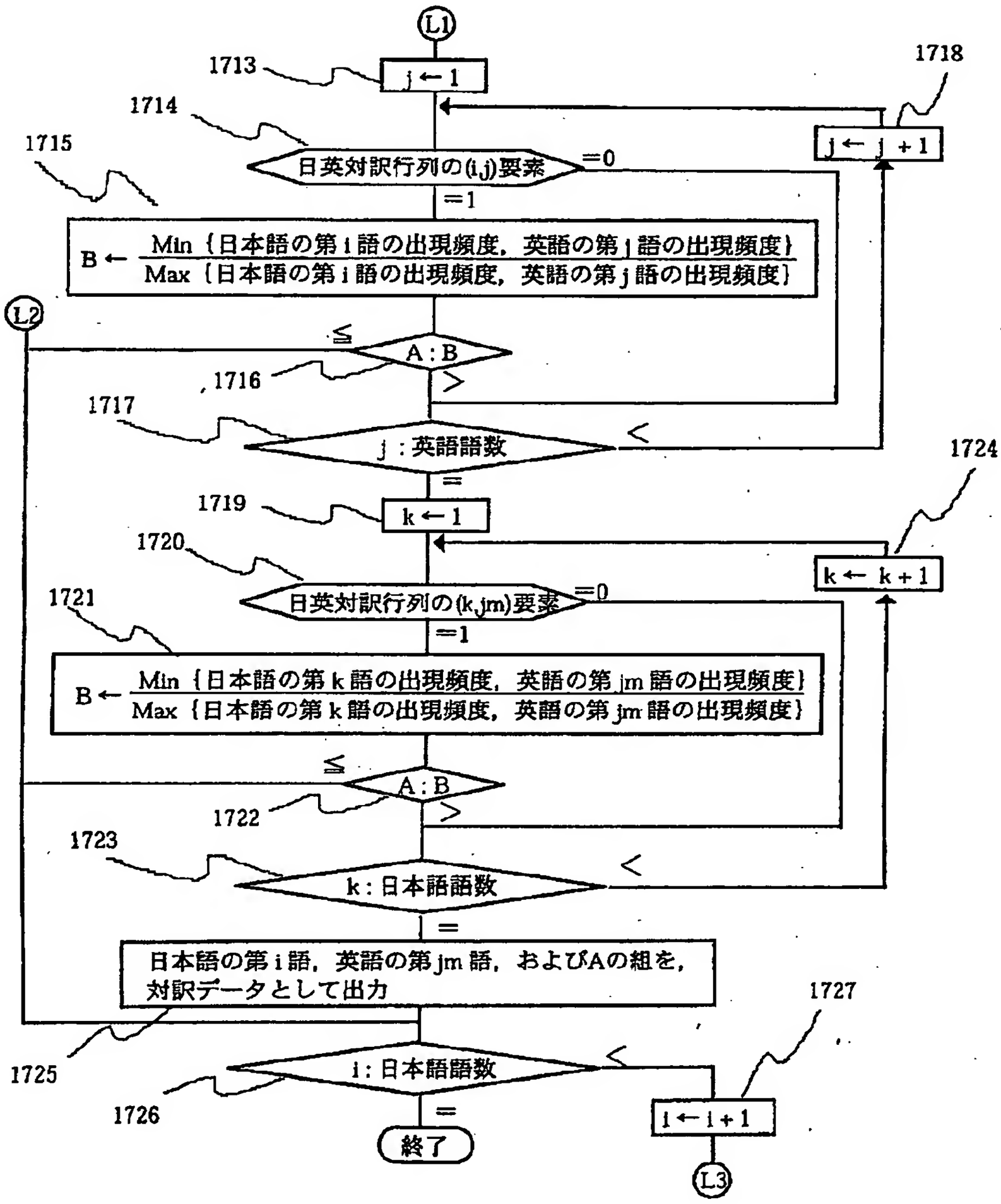
【図１７】

図１７


開始
i ← 1
1701
L3
1703
j ← 1, A ← -1
1702
日英相関行列の(i,j)要素 : A
=
>
1704
A ← 日英相関行列の(i,j)要素,
jm ← j,　N ← 1
N ← N + 1
1705
1707
j ← j + 1
1706
j : 英語語数
<
=
1708
N : 1
>
=
1709
k ← 1
≧
日英相関行列の(k,jm)要素 : A
1712
k ← k + 1
<
L2
1710
k : 日本語語数
<
=
1711
L1

【図１８】

図１８


L1
1713
j ← 1
1714
1715
日英対訳行列の(i,j)要素
=0
=1
1718
j ← j + 1
B ← Min {日本語の第 i 語の出現頻度，英語の第 j 語の出現頻度} / Max {日本語の第 i 語の出現頻度，英語の第 j 語の出現頻度}
L2
≦
A : B
1716
>
1717
j : 英語語数
<
=
1719
k ← 1
1720
日英対訳行列の(k,jm)要素
=0
1721
=1
1724
k ← k + 1
B ← Min {日本語の第 k 語の出現頻度，英語の第 jm 語の出現頻度} / Max {日本語の第 k 語の出現頻度，英語の第 jm 語の出現頻度}
≦
A : B
1722
>
1723
k : 日本語語数
<
=
日本語の第 i 語，英語の第 jm 語，およびAの組を，
対訳データとして出力
1727
1725
i : 日本語語数
1726
=
終了
i ← i + 1
L3